Determinación de Relevancia de Palabras para Procesos con Dominios Restringidos

Germán Rosenbrock[1] – Sebastián Trossero[1] – María Daniela López de Luise[2] – Claudia Alvarez[1] – Andrés Pascal[13] – Fernando Heit[1]
[1]Facultad de Ciencia y Tecnología – Universidad Autónoma de Entre Ríos
[2] CI2S Lab
[3]Facultad Regional Concepción del Uruguay – Universidad Tecnológica Nacional
rosenbrock.german@uader.edu.ar – trossero.sebastian@uader.edu.ar - mdldl@ci2s.com.ar – claudiaalvarez2000@gmail.com - andrespascal2003@yahoo.com.ar - fernandoandresheit@gmail.com

Resumen

En este trabajo se propone un modelo basado en Minería de Textos para la determinación de relevancia que permita la extracción de palabras específicas de un dominio (Domain-Specific Word Extraction). El alcance de la presente propuesta se remite a determinar la importancia de las palabras en el ámbito de regulaciones universitarias, en base a corpus definidos específicamente para evaluar y validar este contexto restringido. Para esto, se emplean cuatro corpus, tres de ellos de dominios relacionados con regulaciones pero aplicados a otros fueros: Regulaciones Universitarias, Regulaciones Impositivas, Regulaciones del Código Civil y un corpus genérico. Se presentan y aplican tests estadísticos pertenecientes a la minería de textos para lenguaje español, y finalmente se comparan las palabras más relevantes del dominio de las regulaciones universitarias con un conjunto de palabras claves extraídas manualmente por especialistas, a fin de validar la propuesta.

Palabras clave: *Minería de textos, procesamiento de lenguaje natural, regulaciones universitarias, extracción de palabras.*

1. Introducción

Muchas ramas de la Inteligencia Computacional aportan al propósito de construir una Inteligencia Artificial desde principios de la década de 1950. Hoy en día, cientos de científicos han contribuido a este campo de muchas maneras, fundando varias ramas que reconocen diferentes tipos de inteligencia [1-3].

Desde la perspectiva del lenguaje existen principalmente aplicaciones relacionadas con el Procesamiento del Lenguaje Natural (PNL). El paso clave es la consideración de diferentes niveles analíticos con diversa complejidad y grados de éxito [4]. Éstas descomponen el problema original aumentando gradualmente el enfoque y el alcance del proceso en analogía con las categorías ya conocidas de la lingüística tradicional: fonología, morfología, sintaxis, semántica y pragmática [5].

Existen entornos de trabajos y plataformas [6-8], ontologías [9], heurísticas [10, 11], lenguajes, corpus, diccionarios e interfaces que forman parte de una gran lista de herramientas útiles con diferente disponibilidad y aplicabilidad. El proyecto actual utiliza su propio corpus, construido a partir de un conjunto de regulaciones universitarias en lenguaje español, la mayoría de ellas escaneadas y convertidas a archivos pdf o archivos de texto [10]. Las heurísticas y las herramientas que son base del proyecto PTAH, al cual pertenece el presente trabajo, se encuentran en Python y WEKA y son principalmente basados en técnicas conocidas como Wavelets Lingüísticos Morfosintácticos (MLW por sus siglas en inglés), Aprendizaje Automático (AA) con Redes Neuronales (NN por sus siglas en inglés) para clasificación, agrupaciones del tipo Maximización de Expectativas (ME), Mapas Autoorganizados de Kohonen (SOM, por sus siglas en inglés) para la auto-indexación de documentos y Sistemas Expertos (ES, por sus siglas en inglés). Para la implementación de la interface con el humano se emplea un lenguaje de marcado del tipo inteligente denominado por sus siglas en inglés como AIML [12]. El prototipo PTAH es un robot de software, comúnmente denominado chatterbot, chatter bot, chat bot, o asistente virtual.

Considerando el corpus como uno de los pasos clave en el flujo de trabajo de chatter bot, el procesamiento de los documentos y su conversión en unidades textuales cobra relevancia ya que determina la precisión con la que interpreta al humano cuando le interpela en el

lenguaje natural. Este artículo se centra en la dificultad de aplicar los modelos de clasificación con palabras claves de dominios genéricos extrapolados a dominios restringidos. La principal contribución de este artículo es la evaluación de la factibilidad de un modelo puramente estadístico para determinar la relevancia de cada palabra dentro de un dominio específico.

La siguiente sección (sección 2) describe el contexto de esta investigación, ya que PTAH necesita no solo administrar un vocabulario de dominio restringido, sino también una restricción en tiempo real para responder las preguntas de los usuarios. La sección 3 describe específicamente el problema de la extracción de palabras claves y la relevancia de las mismas en un contexto de interés. La sección 4 explica el procesamiento de los datos y cómo se aplica el modelo estadístico sobre los mismos. Finalmente, los resultados se muestran en la sección 5, y las conclusiones y trabajo futuro en la sección 6.

2. El Contexto del Chatter bot

El primer chatter bot es Eliza [13], creado por J. Weizenbaum. Aunque pretendía imitar a un psicoanalista rogeriano, fue el inicio de un nuevo tipo de sistemas inteligentes capaces de hablar con humanos en lenguaje natural para diferentes propósitos (también conocidos como chatter bots). Merecen una breve mención las charlas seminales Dr. Colby' Parry [14, 15], y Wallace's Alice (1995) [12, 16]. El enfoque general consiste en definir patrones y plantillas utilizados en las expresiones del lenguaje.

El chatter bot PTAH, puede dialogar con los usuarios sobre temas generales y de regulación académica, lo que constituye un dominio de conceptos restringido. Como parte del proceso de generación de respuestas realiza consultas internas en tiempo real. Utiliza técnicas propias del procesamiento de lenguaje natural para acceder a los documentos de su corpus y responder a las consultas de los usuarios. El proyecto define un modelo de razonamiento lingüístico en el ámbito de las regulaciones. Las leyes y reglamentos se recopilan y segmentan en piezas que se asocian a los metadatos de MLW y se organizan en un SOM [17]. Más detalles sobre el prototipo se encuentran en publicaciones anteriores [11, 17].

3. Relevancia de Palabras

La determinación de la relevancia para la extracción de palabras es utilizada en distintas tareas del Procesamiento del Lenguaje Natural para las traducciones, resúmenes automáticos de textos, agrupamiento de documentos, búsqueda y recuperación de información automática, etc.

Existen varias propuestas que de distintas maneras calculan un puntaje para cada palabra según su frecuencia en un dominio y las penaliza o premia según la rareza de esa palabra en otros dominios [18-21].

La mayoría de los autores plantea comparaciones entre el puntaje dado por el tradicional TF-IDF y otras métricas propuestas como C-Valor, Consenso de dominio, Relevancia de dominio, Entropía Inter Dominio, entre otras.

También es posible la utilización de contextos (embeddings) previamente entrenados, que establecen una forma de representación de las palabras de un documento. Además de representar las palabras aporta información de contexto dentro del documento y de similaridad con otras palabras.

De estas propuestas brevemente explicadas, este trabajo toma y extiende al modelo basado en Entropía Inter Dominio [18].

4. Caso de estudio

Es importante reiterar que en este trabajo se estudia la plausibilidad de un modelo basado en inferencia estadística, con lo que se está implícitamente buscando establecer si es requerido un modelo de mayor sesgo heurístico para cumplir la tarea objetivo de derivación automática de palabras relevantes en dominio restringido. En primera instancia se adapta el método propuesto en [18], que se encuadra en la clasificación pretendida.

Como primer paso se generan cuatro corpus relacionados con regulaciones pero aplicados a fueros distintos: Regulaciones Universitarias, Regulaciones Impositivas, Regulaciones del Código Civil, y un corpus genérico para validar la eficiencia comparada al cambiar de dominio. El primer corpus, correspondiente a resoluciones universitarias, se basa en 27 documentos que forman parte de reglamentación académica. El segundo corpus, consiste en resoluciones de tipo impositivas extraídas del sitio web oficial de la Administración Federal de Ingresos Públicos (AFIP) [22]. Se consideran todas las resoluciones publicadas entre enero de 2021 y agosto de 2022. El tercer corpus, constituido por regulaciones civiles, corresponde a documentos que forman parte del Código Civil y Comercial de la Nación. El último corpus tiene como fin realizar un contraste con los corpus de dominio restringido a regulaciones. Está conformado con documentos de textos genéricos extraídos de los 500 artículos más visitados de Wikipedia en español [23].

Siguiendo la metodología propuesta por Chang, se organizan los corpus cada uno en su propio directorio. De esta forma, ciertas partes del procesamiento se ven facilitados, por ejemplo el acceso con CountVectorizer para leer archivos desde directorios. CountVectorizer es una herramienta de la librería Scikit-learn de Python

que permite convertir una colección de textos o documentos en una matriz de frecuencias de ocurrencias. Este proceso es base del conocido mecanismo de tokenización de documentos (generación de etiquetas especiales).

El corpus en su conjunto alcanza un total de 34.843 palabras distintas. El pre-proceso para realizar la tokenización consiste en pasar a minúsculas las palabras, quitar los números, aplicar *stemming* [24] y eliminar stop words (normalmente palabras cuya alta frecuencia permiten inferir que no contienen significación semántica relevante a ningún tipo de contexto). La eliminación de stop words y el stemming se realizan con la librería NLTK. También fueron excluidas palabras con una ocurrencia menor o igual a 2 documentos. Al finalizar este proceso el data set queda con un total de 17.261 palabras.

La frecuencia de término normalizada para cada palabra en cada uno de los dominios, consiste en sumar todas las ocurrencias del término y dividir por la suma de la ocurrencia de todos los términos. Esta frecuencia es útil para calcular el puntaje de su relevancia relativa dentro del contexto del dominio.

A fin de identificar los términos independientes, que se encuentran distribuidos en todos los dominios, se calcula la Entropía inter-dominio. Las palabras con un valor alto en esta métrica no están estadísticamente asociadas a ningún dominio en particular y por lo tanto se pueden descartar, o bien ponderar con un valor casi nulo en cuanto a su relación con el dominio específico.

El paso final consiste en ordenar las palabras, de mayor a menor, en función del puntaje obtenido para el dominio de interés, que en este caso son las resoluciones regulatorias universitarias.

5. Resultados

En la **Tabla 1** se muestran los resultados obtenidos de calcular la relevancia de las palabras para el dominio específico de las resoluciones académicas. Las palabras se encuentran ordenadas en forma descendente por puntaje, siendo las de mayor puntaje las palabras más relevantes.

Tabla 1. Puntajes de las palabras más relevantes de las resoluciones universitarias.

Palabra	Puntaje
docente	39233
universitaria	27924
educación	22495
cátedra	20378
académica	19065
consejo	19002
alumnado	18915
decano	14959
carrera	10769
asignatura	10691
facultad	9588
concursado	9321
rector	7129
institución	6869
profesor	6289
aspirante	5804
adscripción	5236
consejero	4968
docencia	4772
superior	4692

Considerando las 20 palabras más relevantes obtenidas por el modelo estadístico aplicado en este trabajo, en comparación con las palabras clasificadas manualmente como relevantes, se puede observar que sólo 10 de las 20 son identificadas correctamente como relevantes (ver **Tabla 2**).

Tabla 2. Palabras relevantes por método de selección.

Palabra	SI/NO
docente	SI
universitaria	SI
educación	NO
cátedra	NO
académica	SI
consejo	SI
alumnado	SI
decano	NO
carrera	SI
asignatura	NO
facultad	SI
concursado	SI
rector	NO
institución	NO
profesor	NO
aspirante	NO
adscripción	NO
consejero	NO
docencia	SI
superior	SI

Para contrastar las palabras relevantes del dominio de las resoluciones universitarias contra los demás tipos de resoluciones utilizadas, se presenta una nube de palabras (ver **Figura 1**). Se observan en distintos tonos de grises las palabras generales del dominio de las resoluciones, mientras que con rojo se resaltan las palabras específicas de las resoluciones universitarias o académicas.

Figura 1. Nube de palabras relevantes.

En la misma se puede apreciar que hay palabras importantes en el dominio de las regulaciones o reglamentaciones como por ejemplo *artículo*, *general* o *resoluciones*, que no están resaltadas como relevantes para el dominio de reglamentaciones universitarias, como sí lo están *docente*, *cátedra* o *alumnado*.

Para validar las bondades del modelo estadístico en cuanto a su capacidad de reconocer palabras claves respecto del proceso manual, se establece un punto de corte tomando palabras con puntaje mayor a 1.000. Se obtienen entonces 94 palabras relevantes para el dominio de resoluciones académicas. Comparando estas palabras con las 60 extraídas por especialistas, suponiendo a los mismos como resultado verdadero y confiable, es posible evaluar la sensibilidad en base al porcentaje de palabras que comparten la extracción manual y el modelo estadístico. En la **Tabla 3** se observan las palabras clasificadas manualmente, y cuáles de ellas a su vez son identificadas por el modelo estadístico como palabra relevante.

Tabla 3. Palabras relevantes si fueron o no seleccionadas manualmente.

Palabra	SI/NO	Palabra	SI/NO
académico	SI	ingreso	NO
adjunto	NO	intercambio	NO
alumno	SI	inscripción	NO
área	NO	interino	NO
asociado	NO	investigador	SI
bibliografía	NO	juicio	NO
calendario	NO	jurado	SI
calificación	NO	licencia	NO
cargo	NO	mesa	NO
carrera	SI	nota	NO
correlativa	NO	objetivo	NO
comienza	NO	ordinario	NO
concurso	SI	posgrado	SI
consejo	SI	práctica	NO
convenio	NO	presupuesto	NO
cuatrimestre	NO	promoción	NO
cursar	SI	proyecto	SI
dedicación	NO	regular	NO
designación	SI	rendir	NO
directivo	SI	requisito	NO
docente	SI	sanción	NO
electiva	NO	superior	SI
equivalencia	NO	supervisada	NO
exclusiva	NO	titular	SI
evaluación	NO	universidad	SI
facultad	SI	votar	NO
final	NO	estudiante	NO
grado	NO	concursado	SI
graduado	SI	tribunal	NO
inasistencia	NO	beca	NO

Se observa que el modelo solo logra reconocer 19 palabras de las extraídas manualmente, dando una sensibilidad del 31.67%.

6. Conclusiones y trabajos futuros

Este trabajo pretende determinar la viabilidad de definir por inferencia estadística un modelo capaz de seleccionar automáticamente palabras claves o relevantes para los documentos en un contexto de dominio restringido con acceso en tiempo real para un chat bot.

De los tests realizados el modelo con inferencia estadística logra identificar ciertas palabras que no habían sido identificadas en el proceso de clasificación manual. Como contraparte, no logra reconocer una gran cantidad de palabras importantes para el dominio, llegando a tan solo el 31.67% de las mismas. Dado que la muestra y corpus establecidos en este trabajo son mayormente la recreación del contexto real para el problema establecido, se puede afirmar que la pobreza de los resultados permiten afirmar que los modelos de inferencia estadística presentan una severa limitación. Por lo tanto se hace necesario generar el modelo con inferencia heurísticas de sesgo superior.

En consecuencia, como trabajo futuro se plantea continuar con el desarrollo de otros modelos para el reconocimiento de palabras relevantes en un dominio, pero utilizando técnicas heurísticas.

Adicionalmente se planea ampliar la base de datos de palabras claves del chatter bot, hasta este momento compuesta solamente por las extraídas de forma manual, complementando con las obtenidas como más relevantes en este trabajo, lo que reflejará la variación esperable del corpus mediante el proceso que utiliza el chatter bot para responder a las consultas de los usuarios.

Referencias

[1] Blanes Villatoro A. (2020) “La Teoría de las inteligencias múltiples”. 1º Genética UAB.

[2] Lizano Paniagua K., Umaña Vega M. (2008) “La Teoría de las Inteligencias Múltiples en la Práctica Docente en Educación Preescolar” Revista Electrónica Educare. E-ISSN: 1409-4258.

[3] Howard Gardner (2001) "Estructuras de la mente. La Teoría de Las Inteligencias Múltiples" Fondo de cultura económica. ISBN: 958-38-0063-5

[4] Winograd, T. (1972) "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language". Cognitive

[5] Manning C., Schütze H. (1999) Foundations of Statistical Natural Language Processing. MIT Press. Psychology Vol.3 Nº 1

[6] Corcho, O., López Cima, A., Gómez Pérez, A. (2006) "A Platform for the Development of Semantic Web Portals". ICWE 06. ACM

[7] WEBODE (2022) http://webode.dia.fi.upm.es/WebODEWeb/index.html

[8] Paroubek, P., Schabes, Y., Joshi, A. (1992) "XTAG - A Graphical Workbench for Developing Tree-Adjoining Grammars". Third Conference on Applied Natural Language Processing

[9] Aguado de Cea, G., Álvarez de Mon y Rego, I., Pareja Lora, A. (2002) "Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de la Web Semántica: OntoTag". Iberoamerican Magazine of AI. No. 17. pp. 37 – 49

[10] Vila, K., Díaz, J., Fernández, A., Ferrández, A. (2010) "An Approach for Adding Noise Tolerance to Restricted-Domain Information Retrieval". NLDB 2010. Lecture Notes in Computer Science, vol 6177

[11] López De Luise, D. (2011) "Morphosyntactic Linguistic Wavelets for Knowledge Management". In-Tech Open book “Intelligent Systems”, ISBN 979-953-307-593-7

[12] Wallace, R. (2003) "The Elements of AIML Style". ALICE AI FOUNDATION

[13] Weizenbaum, J. (1966) "ELIZA A Computer Program for the Study of Natural Language Communication between Man and Machine". Communications of Association for Computing Machinery 9, pp. 36-45

[14] Colby K.M., Hilf F.D., Weber S., Kraemer J. (1972) "Turing-Like Indistinguishability Tests for the Validation of a Computer Simulation of Paranoid Processes". A.I., 3, 199-222

[15] Weizenbaum, J. (1976) "Computer power and human reason". San Francisco, CA. W.H. Freeman

[16] ALICEBOT (2022) alicebot.blogspot.com

[17] López De Luise, D., Álvarez, C., Pascal, A., Pancrak, C. (2020) “Chatbots: Autoexpansion Approach to Improve Natural Language Automatic Dialogs”. In Proc. of Int. IEEE ARGENCON 2020

[18] Chang J. (2005) "Domain Specific Word Extraction from Hierarchical Web Documnets: A First Step Toward Building Lexicon Trees from Web Corpora".

[19] Sapan S., Sarath S., Sreedhar R. (2019) “Similarity Driven Unsupervised Learning for Materials Science Terminology Extraction”.

[20] Fedorenko D., Astrakhantsev N., Turdakov D. (2013) “Automatic Recognition of Domain-Specific Terms: an Experimental Evaluation”

[21] Suman D., Radhika M. (2020) “Unsupervised Technical Domain Terms Extraction using Term Extractor”

[22] AFIP (2022) https://www.afip.gob.ar

[23] Wikipedia (2022) https://es.wikipedia.org/wiki/Wikipedia:Ranking_de_visitas

[24] Jivani A. (2011) "A Comparative Study of Stemming Algorithms".